CLewR: Curriculum Learning with Restarts for Machine Translation Preference Learning

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated competitive performance in zero-shot multilingual machine translation (MT). Some follow-up works further improved MT performance via preference optimization, but they leave a key aspect largely underexplored: the order in which data samples are given during training. We address this topic by integrating curriculum learning into various state-of-the-art preference optimization algorithms to boost MT performance. We introduce a novel **c**urriculum **le**arning strategy **w**ith **r**estarts (CLewR), which reiterates easy-to-hard curriculum multiple times during training to effectively mitigate the catastrophic forgetting of easy examples. We demonstrate consistent gains across several model families (Gemma2, Qwen2.5, Llama3.1) and preference optimization techniques. We publicly release our code at `https://link.hidden.4.blind.review`.

1 Introduction

Large language models (LLMs) have enabled zero-shot approaches in multilingual machine translation (MT) (Touvron et al., 2023). Methods for improving the MT abilities of LLMs can be broadly categorized into pre-training and post-training approaches. The former typically employ continual pre-training over large-scale monolingual or high-quality parallel data (Alves et al., 2024; Cui et al., 2025; Xu et al., 2023). In contrast, post-training approaches aim to improve translation quality by employing preference optimization techniques, such as Direct Preference Optimization (DPO) (Rafailov et al., 2023), to distinguish high-quality translations from low-quality ones. Building on this line of work, Xu et al. (2024b) proposed Contrastive Preference Optimization (CPO), a reference-free technique that assesses pair distances based on log-probability differences only. More recently, Xu et al. (2024a) introduced Adaptive Rejection Preference Optimization (ARPO), which further improves CPO by incorporating an adaptive penalty for the unpreferred term.

Despite the significant advances in preference optimization (PO) techniques (Rafailov et al., 2023; Xu et al., 2024a,b), a key factor that can significantly influence performance remains underexplored: the order in which data samples are processed during training. This aspect is central to *curriculum learning* (Bengio et al., 2009), a paradigm that studies how models can learn from easy to hard. The survey of Soviany et al. (2022) explains that an easy-to-hard learning can be created by manipulating distinct factors, namely the data (Chang et al., 2021; Jarca et al., 2024; Nagatsuka et al., 2023), the model (Croitoru et al., 2025b; Sinha et al., 2020) or the target task (Liu et al., 2020a; Narvekar et al., 2016). Organizing the samples in a certain order falls under the umbrella of *data-level curriculum*. In the realm of data-level curriculum, researchers have explored both easy-to-hard and hard-to-easy data organizations, the latter being known as anti-curriculum (Ankner et al., 2024; Florensa et al., 2017; Jarca et al., 2025). Regardless of the data organization, several recent studies showed that curriculum learning can play an important role in various tasks, e.g. natural language inference (Poesina et al., 2024), intent detection (Gong et al., 2021), question answering (Liu et al., 2018), image classification (Liu et al., 2022), model pre-training (Madan et al., 2024; Nagatsuka et al., 2023), etc. Curriculum learning has also been applied in neural machine translation (NMT) (Kocmi and Bojar, 2017; Liu et al., 2020b; Platanios et al., 2019; Zhan et al., 2021), but contributions in this area predate the era of LLMs, making them hard or impossible to adapt to the novel "pre-training then fine-tuning" paradigm. With the emergence of preference optimization techniques applied during the fine-tuning stage (Rafailov et al., 2023; Xu et al., 2024a,b), some recent works (Croitoru et al., 2025a; Pattnaik

et al., 2024) have integrated curriculum learning into DPO. However, these techniques do not explicitly address catastrophic forgetting (Kirkpatrick et al., 2017), a problem that occurs when samples learned at the beginning are forgotten by the model by the end of the training process, eventually degrading performance.

To this end, we propose a novel data-level curriculum learning framework for MT, in which the easy-to-hard training is restarted at every epoch. Our **c**urriculum **le**arning strategy **w**ith **r**estarts (CLewR) is natively designed to mitigate catastrophic forgetting by iterating through all samples in every training epoch. We empirically demonstrate that CLewR leads to consistent performance gains in MT across several state-of-the-art preference optimization methods (DPO, CPO, ARPO) and LLM (Gemma2, Qwen2.5, Llama3.1) families. Our results show that ClewR not only enhances DPO, CPO and ARPO, but also surpasses another competitor based on curriculum learning, namely CurriDPO (Pattnaik et al., 2024).

In summary, our contribution is threefold:

- We propose curriculum learning with restarts (CLewR), a novel method for preference optimization in MT, where the easy-to-hard curriculum is restarted at every epoch to avoid catastrophic forgetting.
- While previous work enhanced DPO with curriculum (Pattnaik et al., 2024), we introduce curriculum learning to newer preference optimization algorithms, namely CPO and ARPO.
- We show that our method outperforms competing curriculum approaches and consistently improves performance across multiple model families (Gemma2, Llama3.1, Qwen2.5) and preference optimization algorithms (DPO, CPO, ARPO).

2 Method

CLewR. We propose a data-level curriculum strategy, named CLewR, which is tailored to preference optimization in MT. We formally present our curriculum strategy in Algorithm 1. Training preference triplets of the form (x, y_w, y_l) are ordered (in step 9) based on a similarity score $s(y_w, y_l)$ between the chosen (winning) y_w and rejected (losing) y_l translations. More precisely, the easiness of a pair of translations is defined as the similarity difference between the preferred and rejected translation, i.e. a high difference corresponds to an easy pair and a low difference to a hard pair. The similarity score is given by the average (step 7) of multiple MT metrics, thus making CLewR suitable for translation: BLEU (Papineni et al., 2002) (step 3), COMET-22 (Rei et al., 2022) (step 4), and METEOR (Banerjee and Lavie, 2005) (step 5) scores. We provide examples of easy and hard preference samples in Appendix A.3. After sorting the triplets, training proceeds over a number of epochs (steps 10-15). At every epoch, the samples are divided into mini-batches (step 11) in the exact order established at step 9, i.e. there is no random shuffling involved. The easy-to-hard data permutation is reused at every epoch, which helps mitigate catastrophic forgetting. Learning is performed via a given PO method (steps 12-13).

Algorithm 1 CLewR Preference Optimization

1: **Input:** initial policy π_θ, training triplets $\{(x_i, y_i^w, y_i^l)\}_{i=1}^N$, learning rate μ.
2: **for** $i = 1$ to N **do**
3: $\quad b \leftarrow \texttt{BLEU}(y_i^w, y_i^l)$
4: $\quad c \leftarrow \texttt{COMET}(y_i^w, y_i^l)$
5: $\quad m \leftarrow \texttt{METEOR}(y_i^w, y_i^l)$
6: $\quad \hat{b}, \hat{c}, \hat{m} \leftarrow \texttt{normalize}_{(0,1)}(b, c, m)$
7: $\quad s_i \leftarrow \frac{1}{3}(\hat{b} + \hat{c} + \hat{m})$
8: **end for**
9: $\mathcal{I} \leftarrow \texttt{argsort}_{\uparrow}\left(\{s_i\}_{i=1}^N\right)$
10: **for** epoch $= 1$ to E **do**
11: $\quad$ **for all** batches $\mathcal{B} \subset \mathcal{I}$ **do**
12: $\quad\quad \mathcal{L}_{\text{PO}} \leftarrow \texttt{loss}\left(x_\mathcal{B}, y_\mathcal{B}^w, y_\mathcal{B}^l, \pi_\theta\right)$
13: $\quad\quad \theta \leftarrow \texttt{optimize}(\theta, \mu, \nabla_\theta \mathcal{L}_{\text{PO}})$
14: $\quad$ **end for**
15: **end for**
16: **Output:** optimized model π_θ

CLewR-z. For ARPO (Xu et al., 2024a), we develop an alternative CLewR implementation called CLewR-z, where the curriculum score s (used in step 7 of Algorithm 1) is derived from the ARPO distance z. The ARPO objective modifies the CPO objective to incorporate an adaptive penalty term τ_θ, which controls the importance of the rejected term y_l:

$$\mathcal{L}_{\text{ARPO}} = -\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\pi_\theta(y_w\,|\,x) - \tau_\theta(y_w, y_l)\,\beta\log\pi_\theta(y_l\,|\,x)\Big) + \log\pi_\theta(y_w\,|\,x)\Big]. \tag{1}$$

The value of τ_θ measures the similarity between

y_w and y_l, ranging from 0 to 1:

$$\tau_\theta(y_w, y_l) = \min(e^{\eta \cdot z_\theta(y_w, y_l)} - 1, 1), \quad (2)$$

where η is a hyperparameter that controls the impact of z_θ, and $z_\theta(y_w, y_l)$ encodes the distance between the chosen and rejected responses by measuring the absolute difference in log-likelihoods:

$$z_\theta(y_w, y_l) = \left| \frac{\log(\pi_\theta(y_w|x))}{|y_w|} - \frac{\log(\pi_\theta(y_l|x))}{|y_l|} \right|. \quad (3)$$

For curriculum learning, we employ $s = -z_\theta$. This version is called CLewR-z.

Enhanced ARPO. We further introduce an enhanced variant of ARPO by using a different distance function $z'_\theta(y_w, y_l)$ that also accounts for distances in the evaluation metric spaces. Specifically, we use:

$$z' = \eta_1 \cdot z_\theta + \eta_2 \cdot z_{\text{BLEU}} + \eta_3 \cdot z_{\text{COMET}}, \quad (4)$$

where z_θ is the original distance. For both metrics, z_{metric} is given by $1 - \frac{\text{metric}}{100}$ to normalize each metric to a $(0, 1)$ interval and have the same monotony as the original z_θ. Two dissimilar predictions result in low BLEU and COMET values, so z_{BLEU} and z_{COMET} will be high. Each z is multiplied by a corresponding scalar η, scaling them to similar intervals. We create multiple versions of enhanced ARPO by modifying the scalars η_1, η_2 and η_3. All ARPO versions are listed in Table 3.

3 Experimental Setup

Dataset. We test on the Flores-200 (Costa-Jussà et al., 2022) dataset. For generic LLMs, we use a group of six languages. For MT-adapted models (GemmaX2), we use a group of three languages, following Cui et al. (2025).

LLM backbones. We consider several candidate LLMs for preference tuning: LLama3.1-8B (Grattafiori et al., 2024), Qwen2.5-7B (Qwen et al., 2025), Gemma2-9B (Team et al., 2024), and GemmaX2-9B (Cui et al., 2025). We also consider X-ALMA (Xu et al., 2024a) as a reference baseline, which is based on Llama2 (Touvron et al., 2023).

Preference optimization baselines. We train the models using three different PO algorithms: Direct Preference Optimization Positive (DPOP) (Pal et al., 2024) (an enhanced version of DPO (Rafailov et al., 2023)), CPO with behavior cloning (Xu et al., 2024b), and ARPO (Xu et al., 2024a). We employ these algorithms with and without our CLewR. For ARPO, we integrate both CLewR and CLewR (z').

Another comparison is between CurriDPO (Pattnaik et al., 2024) and CLewR. For CurriDPO, the data is split into three levels of difficulty (easy, medium and hard).

Models	Test on 6 languages (Romance group)			
	en→xx		xx→en	
	BLEU ↑	COMET ↑	BLEU ↑	COMET ↑
X-ALMA (ARPO)	**39.83**	**89.38**	**42.88**	**89.28**
Gemma2-9B	19.85	61.92	34.98	85.57
+ DPOP	$23.26_{\pm4.23}$	$85.94_{\pm0.59}$	$33.17_{\pm1.47}$	$87.77_{\pm0.02}$
+ CurriDPO	$21.81_{\pm0.26}$	$85.64_{\pm0.34}$	$28.33_{\pm2.71}$	$87.27_{\pm0.20}$
+ CLewR	$22.35_{\pm1.85}$	$86.11_{\pm0.18}$	$31.86_{\pm1.40}$	$87.65_{\pm0.17}$
+ CPO	$33.53_{\pm0.09}$	$87.43_{\pm0.05}$	$36.44_{\pm0.06}$	$87.76_{\pm0.03}$
+ CLewR†	$36.24_{\pm0.03}$	$87.85_{\pm0.01}$	$40.33_{\pm0.23}$	$88.49_{\pm0.04}$
+ ARPO	$35.37_{\pm0.02}$	$88.12_{\pm0.04}$	$37.68_{\pm0.63}$	$88.30_{\pm0.10}$
+ CLewR†	$36.63_{\pm0.05}$	$88.12_{\pm0.01}$	$40.44_{\pm0.06}$	$88.56_{\pm0.02}$
+ CLewR-z†	$36.50_{\pm0.02}$	$88.12_{\pm0.03}$	$39.87_{\pm0.34}$	$88.62_{\pm0.06}$
+ ARPO-z'-V1	$36.35_{\pm0.01}$	$88.40_{\pm0.02}$	$40.32_{\pm0.09}$	$88.62_{\pm0.03}$
+ CLewR†	$\mathbf{37.45}_{\pm0.02}$	$88.40_{\pm0.04}$	$41.28_{\pm0.11}$	$88.78_{\pm0.03}$
+ CLewR-z†	$36.98_{\pm0.03}$	$88.38_{\pm0.04}$	$41.22_{\pm0.08}$	$88.84_{\pm0.02}$
+ ARPO-z'-V2	$36.46_{\pm0.05}$	$\mathbf{88.43}_{\pm0.03}$	$40.10_{\pm0.16}$	$88.58_{\pm0.05}$
+ CLewR†	$\mathbf{37.45}_{\pm0.04}$	$88.37_{\pm0.05}$	$41.09_{\pm0.21}$	$88.75_{\pm0.05}$
+ CLewR-z†	$37.08_{\pm0.02}$	$88.38_{\pm0.01}$	$\mathbf{41.45}_{\pm0.08}$	$\mathbf{88.87}_{\pm0.02}$
Qwen2.5-7B	4.10	45.94	9.59	62.83
+ DPOP	$24.43_{\pm0.88}$	$82.90_{\pm0.24}$	$33.06_{\pm0.76}$	$87.42_{\pm0.12}$
+ CurriDPO	$23.80_{\pm0.26}$	$81.90_{\pm0.30}$	$29.12_{\pm2.58}$	$86.77_{\pm0.39}$
+ CLewR	$23.59_{\pm1.63}$	$82.74_{\pm0.41}$	$30.63_{\pm1.97}$	$87.10_{\pm0.27}$
+ CPO	$27.68_{\pm0.03}$	$83.93_{\pm0.06}$	$31.66_{\pm0.03}$	$86.36_{\pm0.03}$
+ CLewR†	$30.05_{\pm0.02}$	$84.15_{\pm0.06}$	$36.28_{\pm0.10}$	$87.33_{\pm0.05}$
+ ARPO	$30.41_{\pm0.05}$	$85.64_{\pm0.04}$	$32.12_{\pm0.10}$	$86.71_{\pm0.04}$
+ CLewR†	$31.56_{\pm0.10}$	$85.38_{\pm0.07}$	$36.69_{\pm0.27}$	$87.62_{\pm0.10}$
+ CLewR-z†	$31.14_{\pm0.06}$	$85.47_{\pm0.04}$	$34.88_{\pm0.08}$	$87.45_{\pm0.02}$
+ ARPO-z'-V1	$31.08_{\pm0.14}$	$85.67_{\pm0.01}$	$35.12_{\pm0.91}$	$87.37_{\pm0.29}$
+ CLewR†	$32.68_{\pm0.02}$	$85.81_{\pm0.01}$	$\mathbf{37.18}_{\pm0.08}$	$87.60_{\pm0.03}$
+ CLewR-z†	$32.68_{\pm0.05}$	$\mathbf{85.93}_{\pm0.05}$	$36.55_{\pm0.01}$	$87.56_{\pm0.02}$
+ ARPO-z'-V2	$30.95_{\pm0.09}$	$85.67_{\pm0.03}$	$36.40_{\pm0.10}$	$\mathbf{87.75}_{\pm0.02}$
+ CLewR†	$\mathbf{32.70}_{\pm0.03}$	$85.78_{\pm0.01}$	$37.16_{\pm0.06}$	$87.64_{\pm0.03}$
+ CLewR-z	$32.67_{\pm0.04}$	$\mathbf{85.93}_{\pm0.02}$	$36.15_{\pm0.08}$	$87.51_{\pm0.02}$
LLaMA3.1-8B	1.70	46.00	1.84	46.30
+ DPOP	$31.53_{\pm0.11}$	$87.05_{\pm0.08}$	$35.78_{\pm0.46}$	$87.82_{\pm0.01}$
+ CurriDPO	$32.46_{\pm0.36}$	$86.94_{\pm0.06}$	$36.01_{\pm0.50}$	$87.75_{\pm0.05}$
+ CLewR	$30.87_{\pm0.37}$	$86.85_{\pm0.07}$	$35.46_{\pm0.65}$	$87.83_{\pm0.03}$
+ CPO	$1.47_{\pm0.34}$	$54.97_{\pm2.52}$	$10.36_{\pm4.48}$	$74.32_{\pm1.78}$
+ CLewR†	$3.97_{\pm0.99}$	$59.66_{\pm3.21}$	$13.08_{\pm1.93}$	$76.76_{\pm3.90}$
+ ARPO	$33.50_{\pm0.20}$	$87.48_{\pm0.05}$	$35.92_{\pm0.56}$	$87.75_{\pm0.11}$
+ CLewR†	$34.49_{\pm0.05}$	$87.44_{\pm0.06}$	$37.99_{\pm0.07}$	$87.94_{\pm0.04}$
+ CLewR-z†	$34.33_{\pm0.09}$	$87.57_{\pm0.04}$	$36.75_{\pm0.17}$	$87.93_{\pm0.03}$
+ ARPO-z'-V1	$33.68_{\pm0.15}$	$87.44_{\pm0.02}$	$37.82_{\pm0.35}$	$87.91_{\pm0.05}$
+ CLewR†	$35.49_{\pm0.04}$	$87.77_{\pm0.01}$	$\mathbf{39.05}_{\pm0.07}$	$\mathbf{88.22}_{\pm0.01}$
+ CLewR-z†	$35.34_{\pm0.05}$	$87.77_{\pm0.03}$	$38.02_{\pm0.11}$	$88.15_{\pm0.02}$
+ ARPO-z'-V2	$33.96_{\pm0.09}$	$87.49_{\pm0.10}$	$37.47_{\pm0.15}$	$87.84_{\pm0.06}$
+ CLewR†	$\mathbf{35.53}_{\pm0.03}$	$87.77_{\pm0.04}$	$38.89_{\pm0.07}$	$88.17_{\pm0.03}$
+ CLewR-z†	$35.43_{\pm0.01}$	$\mathbf{87.80}_{\pm0.04}$	$37.87_{\pm0.09}$	$88.13_{\pm0.04}$

Table 1: Performance across several base models and preference optimization methods, with and without curriculum learning. Best overall results and per-model best results are shown in bold. † represents statistical significance over the non-curriculum method.

Hyperparameter tuning. We fine-tune LoRA (Hu et al., 2022) adapters on top of the base models, using a rank of 64 for the LoRA matrices, a learning rate of $5 \cdot 10^{-5}$, a warmup ratio of 0.1, and a mini-batch size of 4. All models are optimized with Adam (Kingma and Ba, 2015). Gemma2,

Qwen2.5 and Llama3.1 are each trained for three epochs. Since GemmaX2 is already tuned for MT, it is trained in one epoch. We use default values for other hyperparameters.

Training and evaluation protocol. We focus on translation to and from English for languages from the Romance language group used in X-ALMA (Xu et al., 2024a): Catalan (ca), Galician (gl), Italian (it), Portuguese (pt), Romanian (ro) and Spanish (es). We use the X-ALMA[1] preference data (81K train and 7K validation) and select the triplets from the en $\rightarrow$ xx and xx $\rightarrow$ en language pairs, where xx belongs to the Romance language group. For Gemma2-9B, Qwen2.5-7B and LLama3.1-8B, we use all six languages, while for GemmaX2 we select $\{$it, pt, es$\}$, since the model was not adapted for MT on the other three languages.

To evaluate the models, we compute BLEU and COMET-22 across all considered languages for both en $\rightarrow$ xx and xx $\rightarrow$ en directions. For each setup, we perform three training runs and report the average scores and the corresponding standard deviations across runs. For each run, we select the checkpoint with the best validation performance.

4 Results

We focus on two main comparisons: CLewR vs. CurriDPO, and CLewR vs. non-curriculum. In Table 1, we present MT results with various PO methods, with and without curriculum, applied on generic LLMs (Gemma2, Qwen2.5 and Llama3.1).

CLewR vs. CurriDPO. Across all generic LLMs, DPOP+CLewR improves COMET scores over CurriDPO. The improvements, as measured via paired bootstrap resampling with 10,000 samples and a 95% confidence interval, are statistically significant on 2 out of the 3 models (Gemma2 and Qwen2.5). This demonstrates that our CLewR approach consistently provides benefits beyond existing preference optimization methods, even when applied on top of strong baselines.

CLewR vs. base PO methods. CLewR consistently improves both BLEU and COMET scores across models and preference algorithms. Most improvements are statistically significant according to paired bootstrap resampling with a 95% confidence interval over all sentence pairs. Moreover, our ARPO-z'-V1 and ARPO-z'-V2 variants further improve performance over the other preference methods. This indicates that by incorporating external semantic signals derived from MT metrics into z', our versions of ARPO better adjust preference margins. ARPO-z'-V2 obtains the best performance on average. Additional experiments exploring more z' combinations are presented in Appendix A.2, for en$\rightarrow$ro translations.

[1]XALMA preference dataset

Models	Test on 3 languages (es, it, pt)			
	en→xx		xx→en	
	BLEU ↑	COMET-22 ↑	BLEU ↑	COMET-22 ↑
X-ALMA (ARPO)	37.50	89.00	39.40	88.80
Gemma-X2-9B	37.27	88.85	41.06	88.87
+ DPOP	$38.23_{\pm0.01}$	$89.01_{\pm0.02}$	$\mathbf{41.93}_{\pm0.10}$	$89.09_{\pm0.02}$
+ CurriDPO	$38.01_{\pm0.01}$	$89.04_{\pm0.02}$	$41.60_{\pm0.05}$	$89.08_{\pm0.00}$
+ CLewR	$38.24_{\pm0.16}$	$88.93_{\pm0.01}$	$41.80_{\pm0.07}$	$89.09_{\pm0.02}$
+ CPO	$36.36_{\pm0.04}$	$89.04_{\pm0.00}$	$41.00_{\pm0.09}$	$89.05_{\pm0.01}$
+ CLewR†	$37.69_{\pm0.03}$	$88.87_{\pm0.01}$	$41.45_{\pm0.14}$	$89.04_{\pm0.01}$
+ ARPO	$38.50_{\pm0.03}$	$\mathbf{89.14}_{\pm0.01}$	$40.45_{\pm0.02}$	$89.06_{\pm0.00}$
+ CLewR†	$37.83_{\pm0.01}$	$88.93_{\pm0.01}$	$41.46_{\pm0.02}$	$89.06_{\pm0.00}$
+ CLewR-z	$38.35_{\pm0.01}$	$89.12_{\pm0.00}$	$40.29_{\pm0.02}$	$89.02_{\pm0.01}$
+ ARPO-z'-V1	$38.56_{\pm0.04}$	$89.07_{\pm0.01}$	$41.31_{\pm0.13}$	$89.04_{\pm0.02}$
+ CLewR	$38.08_{\pm0.01}$	$88.98_{\pm0.00}$	$41.72_{\pm0.02}$	$89.09_{\pm0.00}$
+ CLewR-z	$38.26_{\pm0.01}$	$89.09_{\pm0.01}$	$40.87_{\pm0.22}$	$89.02_{\pm0.01}$
+ ARPO-z'-V2	$\mathbf{38.58}_{\pm0.04}$	$89.07_{\pm0.00}$	$41.33_{\pm0.02}$	$89.04_{\pm0.00}$
+ CLewR	$38.06_{\pm0.02}$	$88.99_{\pm0.01}$	$41.83_{\pm0.01}$	$\mathbf{89.10}_{\pm0.01}$
+ CLewR-z	$38.24_{\pm0.03}$	$89.08_{\pm0.01}$	$40.91_{\pm0.25}$	$89.02_{\pm0.03}$

Table 2: Performance of GemmaX2 combined with preference optimization methods, with and without curriculum. The new variations of GemmaX2 surpass both the original model GemmaX2 and X-ALMA. The best results for BLEU and COMET across the directions en→xx and xx→en are bolded. † represents statistical significance over the non-curriculum method.

MT-adapted LLM. In Table 2, we compare the current state-of-the-art LLM for MT, GemmaX2, with PO-based variants and observe that the current SOTA can be further improved by applying preference optimization or curriculum learning. However, these gains are less pronounced than those previously observed in Table 1, and the strongest overall performance for GemmaX2 is obtained using DPOP. This suggests that LLMs optimized for MT may benefit less from more complex preference optimization methods and curriculum learning.

Ablation study. We discuss ablation experiments in Appendix A.2.

5 Conclusion

In this paper, we introduced a novel curriculum learning strategy for preference optimization in MT. We demonstrated consistent gains across several general LLM families. We also showed that the adaptive penalty in ARPO based on modified distances z' derived from MT metrics yields additional performance gains. Moreover, we found that CLewR-z can further improve performance over the original curriculum scoring function.

6 Limitations

Due to computational constraints, we limited our experiments to the 6 languages in the Romance languages group as defined in X-ALMA. More experiments on other language groups and other model families would help strengthen the results.

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Zachary Ankner, Naomi Saphra, Davis Blalock, Jonathan Frankle, and Matthew Leavitt. 2024. Dynamic masking rate schedules for MLM pretraining. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 477–487. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the International Conference on Machine Learning*, pages 41–48.

Ernie Chang, Hui-Syuan Yeh, and Vera Demberg. 2021. Does the order of training samples matter? improving neural data-to-text generation with curriculum learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 727–733.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Nicu Sebe, and Mubarak Shah. 2025a. Curriculum direct preference optimization for diffusion and consistency models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2824–2834.

Florinel-Alin Croitoru, Nicolae-Cătălin Ristea, Radu Tudor Ionescu, and Nicu Sebe. 2025b. Learning rate curriculum. *International Journal of Computer Vision*, 133(1):291–314.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. *arXiv preprint arXiv:2502.02481*.

Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. 2017. Reverse curriculum generation for reinforcement learning. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78, pages 482–495.

Yantao Gong, Cao Liu, Jiazhen Yuan, Fan Yang, Xunliang Cai, Guanglu Wan, Jiansong Chen, Ruiyao Niu, and Houfeng Wang. 2021. Density-based dynamic curriculum learning for intent detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3034–3037.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations*.

Andrei Jarca, Florinel-Alin Croitoru, and Radu Tudor Ionescu. 2024. CBM: Curriculum by Masking. In *Proceedings of the 27th European Conference on Artificial Intelligence*, volume 392, pages 314–321.

Andrei Jarca, Florinel Alin Croitoru, and Radu Tudor Ionescu. 2025. Task-informed anti-curriculum by masking improves downstream performance on text. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2192–2201.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic gradient descent. In *Proceedings of the International Conference on Learning Representations*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum Learning and Minibatch Bucketing in Neural Machine Translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 379–386.

Cao Liu, Shizhu He, Kang Liu, and Jun Zhao. 2018. Curriculum Learning for Natural Answer Generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4223–4229.

Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. 2022. ACPL: Anti-curriculum Pseudo-labelling for Semi-supervised Medical Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20706.

Jinglin Liu, Yi Ren, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020a. Task-level curriculum learning for non-autoregressive neural machine translation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3861–3867.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020b. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436.

Neelu Madan, Nicolae-Cătălin Ristea, Kamal Nasrollahi, Thomas B Moeslund, and Radu Tudor Ionescu. 2024. CL-MAE: Curriculum-Learned Masked Autoencoders. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2492–2502.

Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. 2023. Length-Based Curriculum Learning for Efficient Pre-training of Language Models. *New Generation Computing*, 41(1):109–134.

Sanmit Narvekar, Jivko Sinapov, Matteo Leonetti, and Peter Stone. 2016. Source task creation for curriculum learning. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 566–574.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with DPO-positive. *arXiv preprint arXiv:2402.13228*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pulkit Pattnaik, Rishabh Maheshwary, Kelechi Ogueji, Vikas Yadav, and Sathwik Tejaswi Madhusudhan. 2024. Enhancing alignment using curriculum learning & ranked preferences. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12891–12907.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1162–1172.

Eduard Gabriel Poesina, Cornelia Caragea, and Radu Tudor Ionescu. 2024. A Novel Cartography-Based Curriculum Learning Method Applied on RoNLI: The First Romanian Natural Language Inference Corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 236–253.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585.

Samarth Sinha, Animesh Garg, and Hugo Larochelle. 2020. Curriculum by smoothing. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 21653–21664.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130:1526–1565.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024a. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *arXiv preprint arXiv:2410.03115*.

ARPO Version	η_1	η_2	η_2
ARPO	1.5	0	0
ARPO-z'-V1	0	0	0.5
ARPO-z'-V2	$\frac{0.1}{3}$	$\frac{0.1}{3}$	$\frac{0.5}{3}$
ARPO-z'-V3	0	1.5	0
ARPO-z'-V4	0	0	6
ARPO-z'-V5	0	$\frac{1.5}{2}$	$\frac{6}{2}$
ARPO-z'-V6	$\frac{1.5}{2}$	$\frac{1.5}{2}$	0
ARPO-z'-V7	$\frac{1.5}{2}$	0	$\frac{6}{2}$
ARPO-z'-V8	$\frac{0.1}{2}$	0	$\frac{0.5}{2}$
ARPO-z'-V9	$\frac{1.5}{3}$	$\frac{1.5}{3}$	$\frac{6}{3}$

Table 3: ARPO variants and their hyperparameter configurations.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive Preference Optimization: Pushing the boundaries of LLM performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Runzhe Zhan, Xuebo Liu, Derek F Wong, and Lidia S Chao. 2021. Meta-Curriculum Learning for Domain Adaptation in Neural Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14310–14318.

A Appendix

A.1 Overview of Performance Gains

In Figure 1, we present a detailed analysis on each configuration and its statistical improvements. For each model, the most significant increase was obtained by applying CLewR over CPO. CLewR over DPO degrades performance, suggesting that curriculum benefits from more advanced preference algorithms.

A.2 Ablation Study

In Table 4, we evaluate several z' variants and hyperparameters (η, β). The best variants of ARPO typically involve z_{COMET}. z_{BLEU} helps performance when averaged with z_θ and z_{COMET}. Smaller values for η_1, η_2 and η_3 lead to better performance.

A.3 Easy vs. Hard Examples for Curriculum

In Figure 2, we list several triplets (x, y_w, y_l) that are either easy (low similarity $s(y_w, y_l)$) or hard (high similarity $s(y_w, y_l)$). For the hard examples, we underline the differences between chosen and rejected sentences.

Models	Test on romanian en→ro BLEU ↑	COMET-22 ↑
Gemma2-9B	23.25	68.59
+ ARPO	$38.18_{\pm 0.55}$	$90.54_{\pm 0.02}$
+ CLewR	$38.53_{\pm 0.30}$	$90.58_{\pm 0.01}$
+ CLewR-z	$38.49_{\pm 0.20}$	$90.49_{\pm 0.05}$
+ ARPO $\eta = 0.1$	$38.29_{\pm 0.03}$	$90.63_{\pm 0.01}$
+ CLewR	**39.78**$_{\pm 0.02}$	$90.79_{\pm 0.01}$
+ CLewR-z	$39.10_{\pm 0.03}$	$90.61_{\pm 0.01}$
+ ARPO $\beta = 0.5$	$35.62_{\pm 0.10}$	$90.26_{\pm 0.03}$
+ CLewR	$37.31_{\pm 0.01}$	$90.08_{\pm 0.01}$
+ CLewR-z	$36.96_{\pm 0.07}$	$90.41_{\pm 0.01}$
+ ARPO-z'-V3	$36.98_{\pm 0.22}$	$90.23_{\pm 0.02}$
+ CLewR	$38.43_{\pm 0.01}$	$90.32_{\pm 0.01}$
+ CLewR-z	$38.58_{\pm 0.01}$	$90.53_{\pm 0.01}$
+ ARPO-z'-V4	$37.29_{\pm 0.14}$	$90.34_{\pm 0.05}$
+ CLewR	$38.44_{\pm 0.05}$	$90.46_{\pm 0.02}$
+ CLewR-z	$38.46_{\pm 0.04}$	$90.39_{\pm 0.01}$
+ ARPO-z'-V1	$38.33_{\pm 0.26}$	$90.59_{\pm 0.00}$
+ CLewR	**39.71**$_{\pm 0.06}$	$90.75_{\pm 0.02}$
+ CLewR-z	$39.38_{\pm 0.02}$	$90.73_{\pm 0.04}$
+ ARPO ARPO-z'-V5	$37.31_{\pm 0.00}$	$90.18_{\pm 0.01}$
+ CLewR	$38.99_{\pm 0.02}$	$90.54_{\pm 0.02}$
+ CLewR-z	$38.49_{\pm 0.07}$	$90.43_{\pm 0.02}$
+ ARPO-z'-V6	$37.18_{\pm 0.03}$	$90.18_{\pm 0.00}$
+ CLewR	$38.64_{\pm 0.03}$	$90.52_{\pm 0.01}$
+ CLewR-z	$38.53_{\pm 0.01}$	$90.46_{\pm 0.02}$
+ ARPO-z'-V7	$38.06_{\pm 0.03}$	$90.58_{\pm 0.01}$
+ CLewR	$39.52_{\pm 0.04}$	$90.63_{\pm 0.01}$
+ CLewR-z	$38.70_{\pm 0.01}$	$90.77_{\pm 0.01}$
+ ARPO-z'-V8	$38.23_{\pm 0.01}$	$90.51_{\pm 0.02}$
+ CLewR	$39.59_{\pm 0.07}$	$90.76_{\pm 0.05}$
+ CLewR-z	$39.62_{\pm 0.04}$	**90.86**$_{\pm 0.01}$
+ ARPO-z'-V9	$36.79_{\pm 0.08}$	$90.30_{\pm 0.09}$
+ CLewR	$38.43_{\pm 0.06}$	$90.41_{\pm 0.05}$
+ CLewR-z	$37.98_{\pm 0.15}$	$90.31_{\pm 0.07}$
+ ARPO-z'-V2	$38.16_{\pm 0.11}$	$90.50_{\pm 0.02}$
+ CLewR	$39.66_{\pm 0.09}$	$90.72_{\pm 0.01}$
+ CLewR-z	$39.58_{\pm 0.03}$	**90.86**$_{\pm 0.01}$

Table 4: Ablation study on translations from English to Romanian, experimenting with different hyperparameters and variations of z for ARPO.

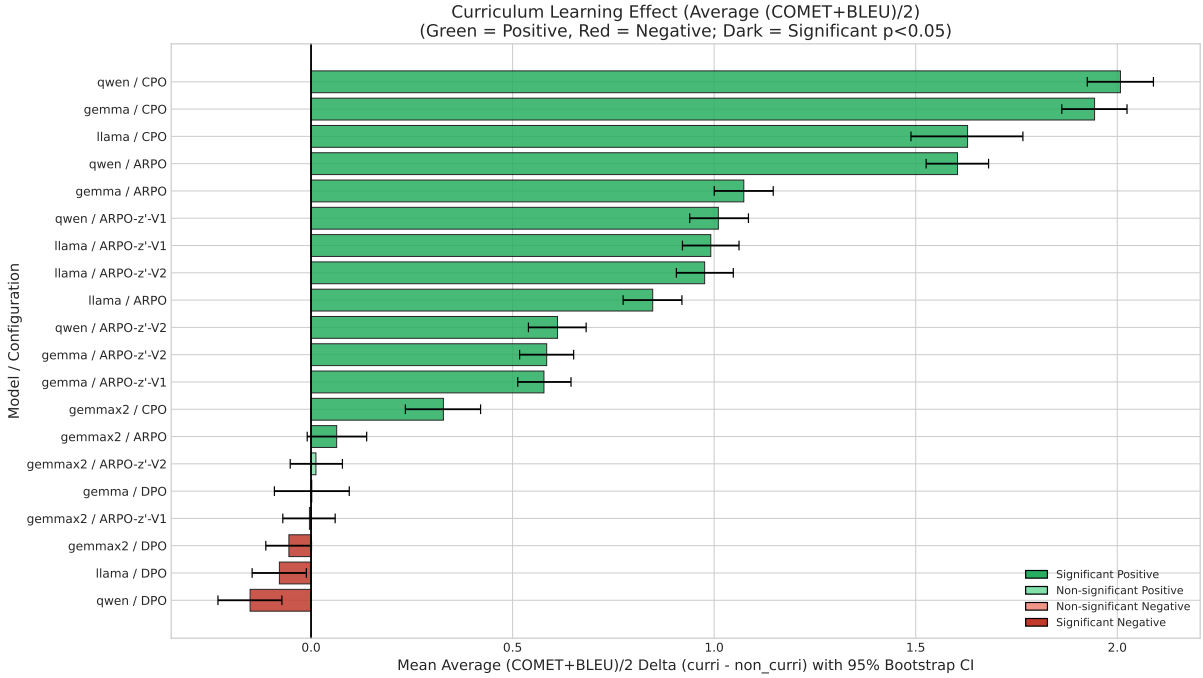

Figure 1: Tested configurations and their improvements from curriculum against the non-curriculum variant. 13 out of 20 configurations show statistically significant improvements, 4 show no significant change, and 3 result in performance degradation. The black line represents the 95% confidence interval.

Examples of training triplets: easy (low similarity between chosen and rejected) and hard (high similarity)

pt→en
EASY: "Para retornar ao nível anterior de energia, seria necessário descarregar a carga extra que foi recebida da iluminação.",
chosen: "To return to their previous energy level, they must get rid of the extra energy they got from the light."
rejected: "To return to the previous level of energy, it would be necessary to discharge the extra charge that was received from the lighting."

es→en
EASY: "Agaché la cabeza y las lágrimas me caían todo el tiempo."
chosen: "I had my head down, and tears were falling the whole time."
rejected: "My head was down and I was crying the whole time."

it→en
HARD: "L'arte della ceramica richiede pazienza e precisione, con artigiani che modellano l'argilla con maestria e la trasformano in pezzi unici e preziosi, ciascuno con la propria storia e carattere."
chosen: "The art of ceramics requires patience and precision, with artisans masterfully shaping the clay into unique and precious pieces, each with its own story and character."
rejected: "The art of ceramics requires patience and precision, with artisans skillfully shaping the clay into unique and treasured pieces, each with its own story and character."

es→en
HARD: "Las nuevas medidas sanitarias han ayudado a controlar la propagación de enfermedades en la comunidad, protegiendo la salud pública y mejorando la calidad de vida."
chosen: "The new health measures have helped control the spread of diseases in the community, safeguard public health, and improve the quality of life."
rejected: "The new health measures have helped control the spread of diseases in the community, protecting public health and improving the quality of life."

Figure 2: Easy vs. hard examples for curriculum learning based on the score s derived from BLEU, METEOR and COMET metrics.